PROJECT 5

Domain and Data

Prepared for the Neural Information Processing Symposium 2003 Feature Extraction Workshop
http://clopinet.com/isabelle/Projects/NIPS2003

MADELON is an artificial dataset, which was part of the NIPS 2003 feature selection challenge. This is a two-class classification problem with continuous input variables. The difficulty is that the problem is multivariate and highly non-linear

MADELON is an artificial dataset containing data points grouped in 32 clusters placed on the vertices of a five dimensional hypercube and randomly labeled +1 or -1. The five dimensions constitute 5 informative features. 15 linear combinations of those features were added to form a set of 20 (redundant) informative features. Based on those 20 features one must separate the examples into the 2 classes (corresponding to the +-1 labels). We added a number of distractor feature called 'probes' having no predictive power. The order of the features and patterns were randomized.

Problem Statement

Our dataset has 499 features to select from. Of those 499 there are 5 relevant features to provide us the highest output. We need to find the most salient features in our model that will get us close to the 5 true features

Solution Statement

Using machine learning technology, we will build algorithm models to find the optimal features. By running these models we will be provided with test scores to support our parameters and algorithms.

Metric

We will continue our process from our first two steps by implementing a pipeline Gridsearch with cross validation along with a SelectKbest using default parameters.

Step 1

We built a Logistic Regression model in a wrapper function using default settings and an L2 penalty. Our benchmark results provided were:

Benchmark Test Score : 0.548484848485
Benchmark Train Score : 0.813432835821

Step 2

Our next step was to continue with the Logistic Regression with default settings and this time using an L1 penalty. The L1 penalty will further punish coefficients to help extract the salient features we are looking for.

RESULTS FROM STEP 2

	coef	features
53	0.057896	feat_053
471	0.045872	feat_471
445	0.040498	feat_445
472	0.036433	feat_472
317	0.035307	feat_317
474	0.034817	feat_474
163	0.033552	feat_163
196	0.030812	feat_196
207	0.030368	feat_207
189	0.029581	feat_189

Step 3

Our last step was to build a pipeline using GridsearchCV along with a SelectKBest and Logistic Regression. Our Pipeline was built using default parameters along with an L2 penalty for the Logistic Regression. Our project best score result was:

grid_search.best_score 0.61499999999999999

Our best estimator results provided the following.

```
Pipeline(steps=[('kbest', SelectKBest(k=1, score_func=<function f_classif at 0x110c19a28>)), ('lr', LogisticRegressio
n(C=1.0, class_weight=None, dual=False, fit_intercept=True,
          intercept_scaling=1, max_iter=100, multi_class='ovr', n_jobs=1,
          penalty='l2', random_state=None, solver='liblinear', tol=0.0001,
          verbose=0, warm_start=False))])
```